This paper was presented at the 19th International Conference on Science and Technology Indicators, September 3-5, 2014, Leiden, The Netherlands.

Atypical combinations are confounded by disciplinary effects

Kevin W. Boyack* and Richard Klavans**

* *kboyack@mapofscience.com*
SciTech Strategies, Inc., Albuquerque, NM, 87122 (USA)

** *rklavans@mapofscience.com*
SciTech Strategies, Inc., Berwyn, PA, 19312 (USA)

Abstract
Uzzi et al. (2013) recently argued that the highest impact articles are likely to reference novel combinations of existing knowledge while still building upon typical combinations. In this study we replicate this intriguing finding using slightly different methods. We also show, however, that the findings are not free from disciplinary effects. For example, physics builds primarily on typical combinations, while multidisciplinary journals participate much more often in atypical combinations. We strongly suspect that atypical co-cited journal combinations, and thus citation rates, are highly dependent on discipline and journal effects.

Introduction
Two new indicators for innovative high impact papers were recently introduced in a *Science* article by Uzzi et al. (2013), hereafter referred to as UMSJ. The authors used co-cited journal-journal relationships to determine whether any pair of cited references is typical or atypical. Using cited references from nearly 18 million articles, they calculated actual and expected counts for each co-cited journal pair, and converted those counts into Z-scores. Negative Z-scores indicate that actual counts are less than expected, and reflect atypical knowledge relationships. Positive Z-scores indicate the opposite – typical knowledge relationships. The authors show that articles that have higher than average typical relationships (using the median Z-score) combined with a high level of atypical relationships (using the left 10th percentile Z-scores) are twice as likely to be highly cited as the average article.

The UMSJ study was designed to test the premise that innovation is often based on original or novel combinations of existing knowledge (Chen et al., 2009; Guimera, Uzzi, Spiro, & Amaral, 2005), while at the same time being strongly based in an existing and well-established paradigm that is robust enough to incorporate new knowledge.

The purpose of this study is to replicate the UMSJ study using a slightly different technique, and to further explore the relationship between novelty (building on atypical knowledge relationships), convention (building on typical knowledge relationships) and citation rates. This paper proceeds as follows. First, we provide detail about the differences between the UMSJ method and our method, and show our replication of their primary results and findings. This is followed a preliminary analysis of disciplinary effects. The paper concludes with a discussion of possible effects from journal impact which may negate their central findings.

Replication
UMSJ calculated Z-scores as $Z = (N_{actual} - N_{expected}) / N_{variance}$ for pairs of co-cited journals where N are journal co-citation counts. Their calculations were based on 17.9 million research articles (1950-2000) from the Web of Science (WoS), and the 302 million references (edges) from these articles to 15,613 cited journals. This formulation gives a negative Z-score to any journal pair where the actual counts are less than the expected counts. Ten Monte Carlo simulations were run that reassigned edges in a random way, while preserving temporal and

distributional characteristics of the original citation network at the paper level. Expected co-citation count values and variances for co-cited journal pairs were calculated from the results of these Monte Carlo simulations.

Using these Z-scores, UMSJ then calculated 10th percentile (left tail) and median Z-scores for each article after ordering the Z scores corresponding to their co-cited journal pairs from lowest to highest. The resulting cumulative probability distributions showed that half of WoS articles had a median Z-score greater than 64, while 41% of those articles had a 10th percentile Z-score that was negative. These two statistics were used as the basis for two indicators. The median Z-score for an article was used to signal conventionality; articles with a median score of greater than the overall median were designated as "high convention". The 10th percentile Z-score was used to signal novelty; articles with a negative 10th percentile Z-score were designated as "high novelty". Upon testing the top 5% highly cited articles (by year), UMSJ found that articles with high convention and high novelty are twice as likely to be highly cited as the average article. Although UMSJ also tested different definitions of novelty (e.g., 1%, 10%) and explored the effect of authorship structure on their results, those additional experiments did not change the overall results. Thus, our study focuses on replicating the primary typical vs. atypical distributions and indicators of convention and novelty that are the basis for the findings of the UMSJ study.

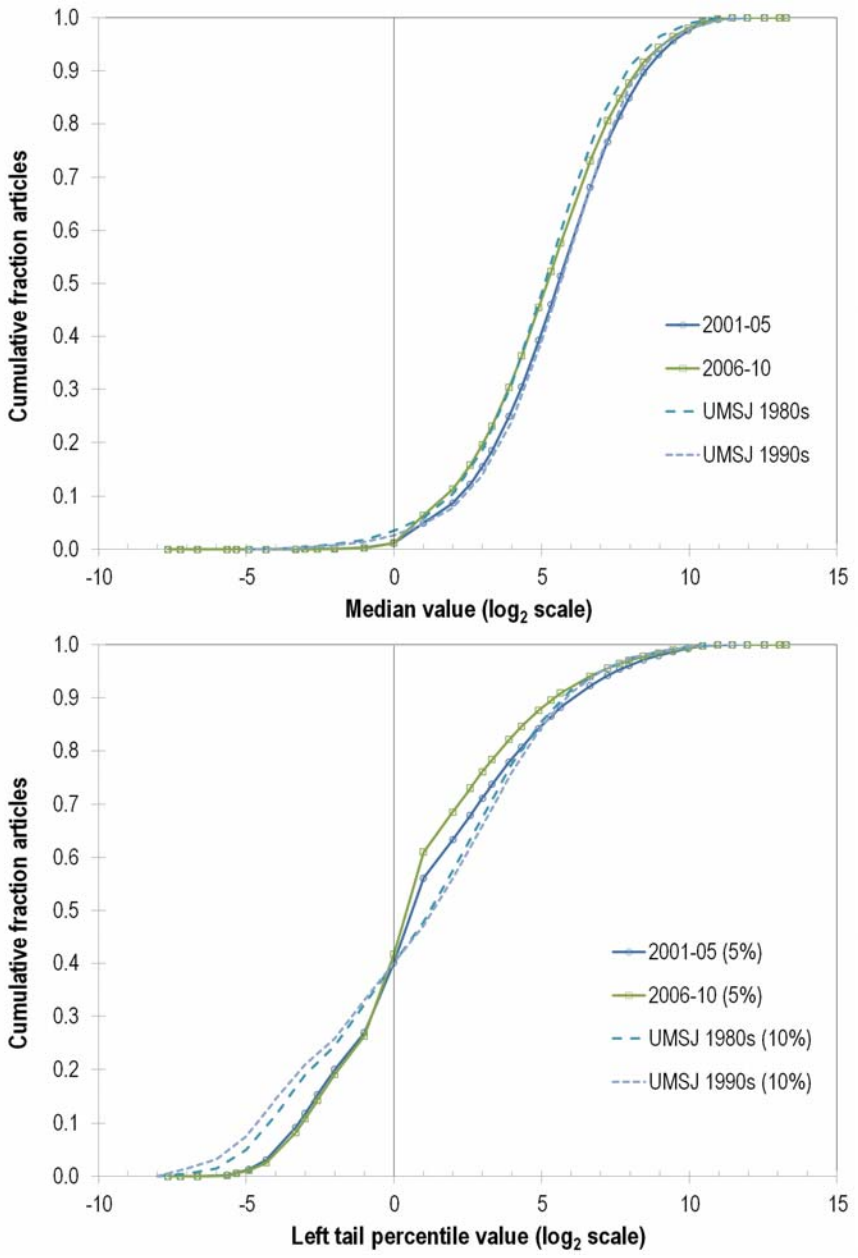
The methodology we used to replicate the UMSJ study differs from theirs in several respects. First, we used Scopus data rather than WoS data. Second, we used a more current ten year dataset (2001-2010) rather than the historical 50 year dataset (1950-2000) used by UMSJ. Our dataset is thus smaller than the one used by UMSJ (12.0M articles + 226M references vs. 17.9M articles + 302M references), but certainly still large enough to provide for valid results. The difference in time window is not expected to be an issue since UMSJ showed results that were comparable for multiple time periods. Third, while UMSJ used articles only, we used articles and conference papers. Scopus indexes much more conference material than does WoS, and since articles and conference papers are both aimed primarily at reporting original research we felt justified in including both document types. Finally, we used a different formulation to calculate typical and atypical relationships. Rather than using Z-scores and Monte Carlo simulations, we calculated K50 statistics for co-cited journal pairs (Klavans & Boyack, 2006). K50 has the same general formulation as the UMSJ Z-scores, $(N_{actual} - N_{expected}) / normalization$. The difference is that the expected and normalization values for K50 are calculated using the row and column sums from the square co-citation count matrix rather than using a Monte Carlo technique. This difference leads to a savings in computation – calculating row and column sums is much less expensive computationally than using multiple Monte Carlo runs. Our K50 distributions are very similar to Uzzi's Z-score distributions, thus suggesting that the additional computation required by multiple Monte Carlo calculations may be unnecessary.

Distributions

Figure 1 compares the distributional characteristics of median and left tail percentile statistics from our study with those of UMSJ. Z-score curves were obtained by transcribing data from Figures 1B, C of Uzzi et al. (2013). Our K50 values have been scaled (multiplied by 10^4) to fall within the same range as the UMSJ Z-scores. Figure 1a shows that while the fraction of papers with negative median K50 values is lower than the UMSJ values, the K50 curves fall between the two UMSJ curves over most of the range. Thus, use of median statistics to designate articles as "high convention" should work similarly with K50 values as it does for the UMSJ Z-scores. For the left tail values, we found that only 30% of articles had a 10th

percentile K50 value that was negative, while 40% of articles had a 5th percentile K50 value that was negative. Figure 1b compares K50 values at the 5th percentile with UMSJ 10th percentile values, and shows that the K50 curves are very similar to the UMSJ Z-score curves. Thus, our use of 5th percentile K50 statistics to designate articles as "high novelty" should perform similarly to the UMSJ 10th percentile Z-scores.

Figure 1. Comparison of median and left tail distributions from K50 statistics with the same distributions based on UMSJ Z-scores.



The K50 distributions are remarkably similar to the UMSJ distributions given that we used a different database, a different metric, and included conference papers along with articles in our calculations. Based on this similarity between distributions, both in the principles behind

their calculation and in practice, replication of additional results from UMSJ using K50 statistics is justified.

Indicators

UMSJ proposed a method for identifying "hit" papers using the principles of novelty and conventionality based on Z-scores and their distributions. To test this method, a 2x2 categorization based on median and 10th percentile Z-scores was used to classify the top 5% highly cited papers (citation counts as of 8 years after publication). We followed the same procedure with some differences. We computed all citation counts to papers as of 2011; thus papers published in 2001 had a ten year citation window while papers published in 2005 had only a 6 year window in which to accrue citations. Also, we used the 5th percentile (rather than the 10th percentile) K50 score as the basis for distinguishing between high novelty and low novelty.

As with UMSJ, our analysis was limited to the top 5% highly cited articles by year. Despite the differences in our test samples, we get 2x2 matrix probabilities that are similar to UMSJ (see Table 1). The differentiation between our high/high (N+C+) and low/low (N-C-) pairs is even higher than that obtained by UMSJ. In addition, the fraction of articles that end up in the N+C+ bin is slightly higher using our method (9.5% vs. 6.7%), suggesting that our calculations can identify even more highly cited papers than can the UMSJ method. Note that the N+C- bin also has a probability of greater than 5% (0.0659), which suggests that novelty plays a greater role than conventionality in the formulation of a "hit" or highly cited article.

Table 1. Probabilities of "hit" papers (top 5% highly cited).

	UMSJ (1990-2000)		**This study (2001-2005)**	
	% sample	**Prob**	**% sample**	**Prob**
High Novelty, High Convention (N+C+)	6.7%	0.0911	9.5%	0.0959
High Novelty, Low Convention (N+C-)	26%	0.0533	30.6%	0.0659
Low Novelty, High Convention (N-C+)	44%	0.0582	40.5%	0.0433
Low Novelty, Low Convention (N-C-)	23%	0.0205	19.4%	0.0205

In summary, we have replicated the distributions and hit paper probabilities introduced in Uzzi et al. (2013) to a high degree, despite differences in methodology. This replication suggests that our process is sufficiently accurate to be used to more deeply explore the relationships between novelty, convention, and citation rates.

Disciplinary effects

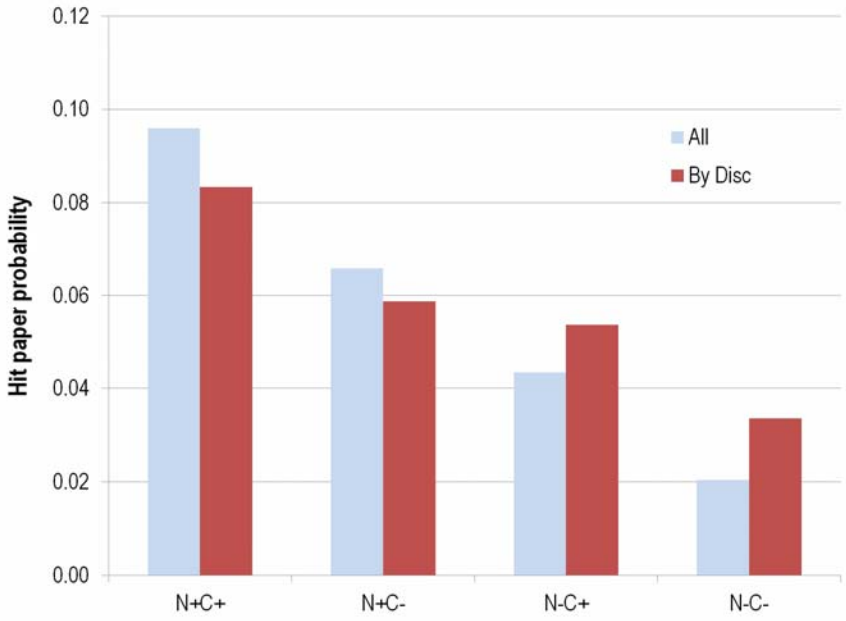
As mentioned above, UMSJ tested multiple definitions of novelty and explored the effect of authorship structure on their results. They also explored the effect of disciplines on their results by examining central tendencies for median and 10th percentile statistics by WoS subject category. They looked at the relationships between novelty, convention, and hit papers for each category, and found that the overall relationships generally held true. However, their detailed results showed that the N+C+ bin in the 2x2 matrix had the highest probability of containing a hit paper for only 64.4% of 243 WoS subject categories. Although this is consistent with the main result on the whole, the fact that this number is not close to 100% suggests that their method is not free from disciplinary effects. It is also well known that impact by discipline is nonlinearly related to size (Katz, 1999, 2000). Thus, we felt it prudent to more deeply explore potential disciplinary effects on the indicators proposed by UMSJ.

Discipline-based sampling

The first, and simplest, test was to calculate 2x2 matrix probabilities using the top 5% highly cited articles where the top 5% was sampled by discipline rather than over the entire sample. We expected different results because the top 5% sample over all disciplines used by UMSJ is naturally enriched in papers from disciplines with high citation rates (e.g., biochemistry, physics) and depleted in papers from disciplines with lower citation rates (e.g., social sciences, engineering). Sampling by discipline will introduce papers with smaller numbers of citations from these lower cited disciplines into our sample at the expense of more highly cited papers from highly cited disciplines.

We took the top 5% of highly cited papers by discipline using the article-based (as opposed to journal-based) discipline-level structure introduced in Boyack and Klavans (2014) and calculated 2x2 matrix probabilities. Figure 2 shows that while discipline-based sampling preserves the probability ordering of bins (i.e., N+C+ highest, N-C- lowest), the separation between the highest and lowest probabilities is much less than for the non-discipline based case. This degradation suggests that the higher probability associated with the non-discipline based case is due to the enrichment of that sample with articles from highly cited disciplines, and is evidence of a larger disciplinary effect than is acknowledged by Uzzi et al. (2013). This does not detract from the fact that, even when disciplines are considered, the combination of typical and atypical combinations associated with these indicators leads to a higher than average incidence of highly cited papers. However, when disciplines are considered the effect is less prominent.

Figure 2. Effect of sampling the top 5% highly cited papers by discipline on probabilities of hit papers based on novelty and conventionality indicators.

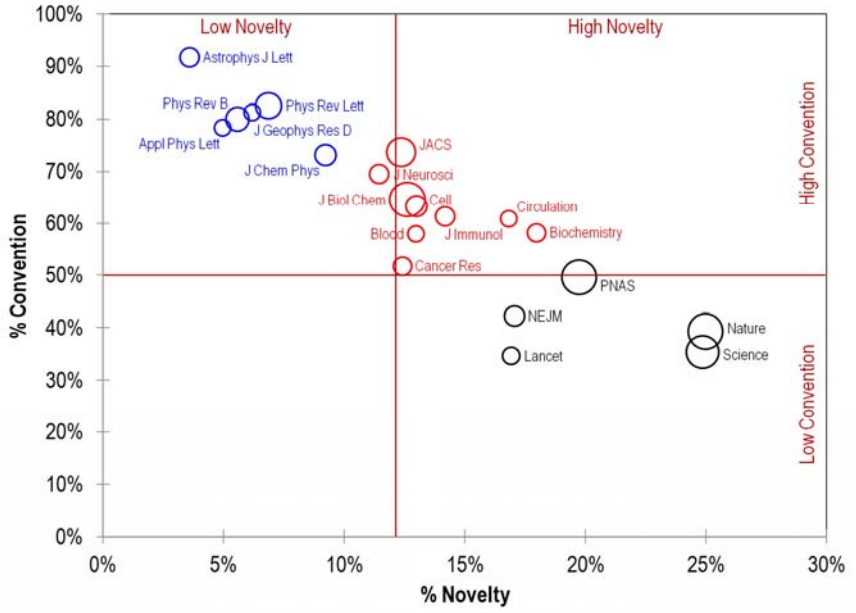


Top 20 knowledge areas

Since UMSJ used journals as surrogates for knowledge areas, we also used journals as the base unit of analysis in our replications of their results. As with disciplines, journals vary widely in size and influence. Thus, we decided to take a closer look at those journals that contributed most to the system of knowledge interactions.

A total of 58,020 separate Scopus journal identifiers were cited by the 12 million articles in our dataset. Although this number seems much larger than the 15,613 journals analyzed by UMSJ, the signal is highly concentrated in a much smaller number of journals. The top 300 journals account for half of the total number of co-citations in the system, while the top 15,600 journals account for 99.6% of the total number of co-citations. Thus, the existence of a long tail in our data has almost no effect on the overall system. We limited our analysis to the top 20 journals, which participated in 15.9% of the co-citations in the system. Four of these journals (*J Biol Chem*, *Nature*, *Science*, and *PNAS*) each participated in more than 1.5% of the total co-citations.

Figure 3. Top 20 co-cited journals plotted as a function of novelty and convention. Circle sizes reflect numbers of co-citations.



Percentages of novel and conventional K50s were calculated for each journal, where %Novel is the fraction of negative K50s, and %Convention is the fraction of K50s above the median (0.00421226) for the entire system. Figure 3 shows these 20 journals, each plotted as a function of their %Novel and %Convention values. The figure has been divided into four quadrants that correspond to the four groupings in the 2x2 matrix mentioned earlier. The dividing line for novelty is at 12.16%, which is the fraction of all co-citations across the system with negative K50 values. Among these 20 journals, three groups can be easily distinguished. Six journals, all of which are highly related to physics, are closely grouped in the low novelty, high convention quadrant (upper left). Nine journals, all of which are related to biochemistry or medicine, are grouped in or very near the high novelty, high convention quadrant (upper right). The remaining five journals are all in the high novelty, low convention quadrant. Three of these journals are clearly multidisciplinary while the other two (*NEJM*, *Lancet*) are broad medical journals, and thus more multidisciplinary than other medical journals. The type of knowledge relationships associated with these prominent journals clearly varies by discipline. Physics is highly associated with typical relationships. Biochemistry and medicine are associated with the pair of relationships promoted by UMSJ –

a combination of typical and atypical relationships. Multidisciplinary journals are more highly associated with atypical knowledge relationships.

We note that this analysis accounts for only 15.9% of the co-citations in the system, and only applies to a few of the top cited disciplines in science. A detailed investigation of the rest of the system may show different effects. Nevertheless, the fact that a large disciplinary effect is seen in the top few journals (which comprise a significant fraction of the overall signal representing typical and atypical relationships) suggests that discipline may be a significant confounding effect as regards these relationships.

Summary

We have replicated the distributions and hit paper probabilities from UMSJ using a slightly different methodology. This replication allows us to proceed to more deeply explore how the notions of novelty and convention might be measured using citation data and our metrics.

The analysis of disciplinary effects above is preliminary; a much more detailed analysis is needed. In addition, the fact that three high impact multidisciplinary journals (*Nature*, *Science*, *PNAS*) account for 9.4% of all of atypical combinations (negative K50 values) suggests that there may be significant journal-level effects as well.

The idea that measurement of novelty might lead to a paper-level indicator of impact type has been intriguing to us for some time (Klavans & Boyack, 2013). While we point out some potential problems with specifics of the UMSJ study, we believe that their underlying logic – that of creating an indicator based on the notion of novelty and distribution tails – is sound. What remains is to identify and test other potential measurements of novelty that are relatively independent of discipline and journal effects.

References

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology, DOI: 10.1002/asi.22990*.

Chen, C., Chen, Y., Horowitz, M., Hou, H., Liu, Z., & Pelligrino, D. (2009). Towards an explanatory and computational theory of scientific discovery. *Journal of Informetrics, 3*, 191-209.

Guimera, R., Uzzi, B., Spiro, J., & Amaral, L. A. N. (2005). Team assembly mechanisms determine collaboration network structure and team performance. *Science, 308*, 697-702.

Katz, J. S. (1999). The self-similar science system. *Research Policy, 28*, 501-517.

Katz, J. S. (2000). Scale-independent indicators and research evaluation. *Science and Public Policy, 27*(1), 23-36.

Klavans, R., & Boyack, K. W. (2006). Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology, 57*(2), 251-263.

Klavans, R., & Boyack, K. W. (2013). *Towards the development of an article-level indicator of conformity, innovation and deviation*. Paper presented at the 18th International Conference on Science and Technology Indicators.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science, 342*, 468-472.